Measuring Robustness with First Relevant Score in the TREC 2012 Microblog Track

Stephen Tomlinson
OpenText
Ottawa, Ontario, Canada
stomlins@opentext.com
http://www.opentext.com/

February 3, 2013

Abstract

In this paper, we measure the effectiveness of various experimental search techniques not just with traditional TREC ad hoc search measures such as Average Precision, R-precision and Precision@30, but also with robust measures based on just the rank of the first relevant item retrieved such as First Relevant Score and Generalized Success@30. We report the results of our experiments conducted in the context of the Real-Time Adhoc Search Task of the TREC 2012 Microblog Track which investigated the effectiveness of ad hoc search of a collection of more than 10 million tweets. For the experimental technique of favoring tweets with urls, we found that both the traditional and robust measures indicated statistically significant increases in the mean score. However, for an experimental blind feedback technique, a technique known to be non-robust as it typically makes poor results even worse, the traditional Average Precision measure indicated a statistically significant increase in the mean score, but some of the measures just based on the rank of the first relevant item successfully discerned a statistically significant decrease in the mean score from the non-robust technique.

1 Introduction

OpenText Search Server®, eDOCS Edition (formerly known as Open Text eDOCS SearchServer™) is a toolkit for developing enterprise search and retrieval applications. The eDOCS SearchServer kernel is also embedded in various components of the OpenText eDOCS Suite[1].

The eDOCS SearchServer kernel works in Unicode internally [2] and supports most of the world's major character sets and languages. The major conferences in text retrieval experimentation (TREC [5], CLEF [1] and NTCIR [3]) have provided judged test collections for objective experimentation with the SearchServer kernel in more than a dozen languages.

This paper describes experimental work with the eDOCS SearchServer kernel (experimental post-6.0 builds) conducted in part by participating in the Real-time Adhoc Search Task of the TREC 2012 Microblog Track.

2 Real-Time Adhoc Search Task

The Real-Time Adhoc Search Task of the TREC 2012 Microblog Track was, given a short query at a particular time, to produce a ranked list of the most relevant tweets in the "Tweets2011" corpus prior to

[1]OpenText, Open Text eDOCS SearchServer and Open Text eDOCS Suite are trademarks or registered trademarks of Open Text Corporation in the United States of America, Canada, the European Union and/or other countries. This list of trademarks is not exhaustive. Other trademarks, registered trademarks, product names, company names, brands and service names mentioned herein are property of Open Text Corporation or other respective owners.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
NOV 2012		**00-00-2012 to 00-00-2012**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Measuring Robustness with First Relevant Score in the TREC 2012 Microblog Track	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **OpenText,10 Rideau Street, 6th Floor,Ottawa, Ontario, Canada,**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Presented at the Twenty-First Text REtrieval Conference (TREC 2012) held in Gaithersburg, Maryland, November 6-9, 2012. The conference was co-sponsored by the National Institute of Standards and Technology (NIST) the Defense Advanced Research Projects Agency (DARPA) and the Advanced Research and Development Activity (ARDA). U.S. Government or Federal Rights License

14. ABSTRACT
In this paper, we measure the e ectiveness of various experimental search techniques not just with traditional TREC ad hoc search measures such as Average Precision, R-precision and Precision@30, but also with robust measures based on just the rank of the rst relevant item retrieved such as First Relevant Score and Generalized Success@30. We report the results of our experiments conducted in the context of the Real-Time Adhoc Search Task of the TREC 2012 Microblog Track which investigated the e ectiveness of ad hoc search of a collection of more than 10 million tweets. For the experimental technique of favoring tweets with urls, we found that both the traditional and robust measures indicated statistically signi cant increases in the mean score. However, for an experimental blind feedback technique, a technique known to be non-robust as it typically makes poor results even worse, the traditional Average Precision measure indicated a statistically signi cant increase in the mean score, but some of the measures just based on the rank of the rst relevant item successfully discerned a statistically signi cant decrease in the mean score from the non-robust technique.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**10**	

that time.

2.1 Downloading the Tweets2011 Corpus

While the Tweets2011 corpus ostensibly consisted of 16,141,812 tweets (sampled from Twitter during the 2-week period of Jan 24-Feb 8, 2011), only 10,624,958 of the tweets were eligible this year (based on the number that NIST successfully downloaded with http status 200 in their 2012 crawl, as listed in the "id-status.01-May-2012" file that NIST provided).

Furthermore, each participant had to download the tweets themselves from twitter.com, one at a time, based on the given list of 16,141,812 urls. The availability of tweets could change over time (such as from users deleting tweets, or a timeout occurring while attempting to download a tweet), hence each participant likely had a different subset of the corpus.

While many of the track participants had downloaded the corpus last year, we had not, and we did not decide to participate in the Microblog track this year until shortly after May 10, 2012 when the Legal Track suddenly announced it was not running this year.

There were many adventures attempting to download the collection which were documented on the track mailing list at the time. Apparently the html pages on twitter.com became much larger (approx May 2012) compared to last year, to approx 89,000 bytes per html page (with just 1 tweet, famously known to be at most 140 characters, scraped per page), which caused the provided download tool to fail from a writeUTF() 64KB limit, and increased the download footprint over the wire to approximately a terabyte. A track participant, Myle Ott, was particularly heroic at fixing the open-source download tool to deal with the html format changes, and also reducing its memory usage, reducing its disk space footprint while increasing the richness of the downloaded information, increasing its speed, and adding the capability to follow redirected tweets.

We performed our download from June 17-29, 2012, using Myle Ott's enhanced crawler "AsyncEmbeddedJsonStatusBlockCrawler" of myleott-twitter-corpus-tools-c254443.zip from https://github.com/myleott/twitter-corpus-tools . The downloading was just done during off hours (overnight and weekends).

We downloaded the urls in random order so any glitches during the download process (e.g., web server rejecting requests for a few minutes) should be spread over the collection rather than clustered to particular tweet time periods.

In our first download pass (June 17-23, 2012), we used the crawler's -noFollow option (which did not follow http status 301 redirects), and it downloaded 11,218,704 tweets successfully, including 10,296,356 of the previously mentioned eligible 10,624,958 tweets (though it was not announced until July 3, 2012 that the eligible list was just the 10,624,958 tweets of status 200 in the NIST download, so we thought we might be missing more of the eligible tweets).

In our second (and final) pass (June 24-29, 2012), we re-attempted to download the 4,923,108 tweets missing after the first pass, and we did not use the -noFollow option (so that the crawler would follow http status 301 redirects, which again, we did not know were ineligible until July 3, 2012).

We ended up downloading 13,551,650 tweets in total, including 10,414,471 of the previously mentioned eligible 10,624,958 tweets (98.0%). (Also, compared to the NIST id-status.01-May-2012 list, we successfully downloaded 2,198,005 of the 3,015,117 tweets it labelled status 301 (redirected), 787,723 of the 815,794 tweets it labelled status 302 (retweet), 130,326 of the 817,273 tweets it labelled status 403 (forbidden), 21,125 of the 868,667 tweets it labelled status 404 (not found), and 0 of the 3 tweets it did not label.)

Cross-checking against last year's relevance assessments (qrels), our download included 94% of the tweets judged relevant in at least 1 topic (2778 of 2965), including 94% of the tweets judged highly relevant in at least 1 topic (530 of 561). Hence we had ample coverage of last year's relevance assessments for doing training experiments.

As a later followup, when the relevance assessments (qrels) for the 2012 task were received on August 31, 2012, it turned out that our download included 99.49% of the tweets judged relevant in at least topic (6233 of 6265), including 99.53% of the tweets judged highly relevant in at least topic (2549 of 2561).

2.2 Indexing

We indexed the 13,551,650 downloaded tweets as follows.

First, we reformatted the tweets to the XML format of the IIT CDIP collection used in the Legal Track in 2006-2009 so that we could re-use essentially the same indexing scripts.

While the downloaded tweets in .json format had a lot of detailed information, e.g., how many followers the author had, the only information we wrote out in our XML version was the tweet itself (e.g., "BBC News - BBC World Service cuts to be outlined to staff http://www.bbc.co.uk/news/entertainment-arts-12283356"), the tweet author (e.g., "ashstreetltd") and the tweet id (e.g., "30198105513140224").

In particular, for each tweet, we added a "<record>" tag at the beginning, followed by the tweet id inside "<tid>..</tid>" tags (a 17-digit number ranging from 28965131362770944 to 35123253429284864), followed by the first 2 digits of the tweet id inside "<DD>..</DD>" tags (a 2-digit number ranging from 28 to 35), followed by the tweet author userid inside "<BX>..</BX>" tags, followed by the tweet (which seemed to already be in XML-safe form, i.e., it used < and > instead of < and >, though it did not escape the ampersand to &, but the ampersand handling seemed unlikely to matter for our purpose), followed by a closing "</record>" tag. We ended up with 212 .xml files totalling 2,123,389,692 bytes.

Our indexing script, for each record, indexed from the "</tid>" tag to the "</record>" tag. Any tags themselves were indexed (we just didn't bother to discard them; a minor side effect is that this meant the term "record" matched every document). Entities (e.g. ">") were converted back to the character they represented (e.g. ">").

This year we used an English stopword list of 86 words to not index (e.g., "the", "of", etc.). We indexed 4 punctuation characters as 1-character words (#, $, %, @). The apostrophe was considered a word separator. The index supported both searching on just the surface forms of the words and also searching on inflections from English lexical stemming. The documents were assumed to be in the UTF-8 character set. Words were normalized to upper-case and any accents were dropped.

2.3 Searching

The track organizers created 60 test topics numbered 51 to 110, each including a short query and the querytweettime. Participants could submit up to 4 runs listing the top-10,000 tweets scored by relevance for each topic. Submissions were due July 10, 2012. The experimental techniques used for our 4 submitted runs (plus one other unsubmitted run produced at the time) are described below.

For reference, here are the codes used in the run names (explained in more detail below): 'i' indicates "uses full-collection idf"; 'h' indicates "added HTTP to query"; 'e0' indicates "uses blind feedback based on first-30 (non-future) retrieved"; 'e' indicates "50% blind feedback, 50% ih".

2.3.1 Baseline Run - otM12i

The submitted run 'otM12i' was produced as follows.

The run just searched for tweets matching a Boolean-OR of the query words or their English inflections by passing the query to the SearchServer IS_ABOUT predicate.

The SearchServer '2:3' relevance method was the same as described in previous years [10]. The relevance function dampened the term frequency and adjusted for document length in a manner similar to Okapi [4] and dampened the inverse document frequency using an approximation of the logarithm. (SET RELEVANCE_DLEN_IMP 750 was used for document length importance.) In runs which used inflectional matching (which was the case for all the runs this year), these calculations were based on the stems of the terms.

To remove at least some of the tweets from after the "querytweettime" given in the topic, a Boolean-AND was also in the query (with relevance weight 0) based on the DOCDATE field (which corresponded to the "DD" field in the XML format described earlier).

For example, for topic 71, for which the query was "Australian Open Djokovic vs. Murray" and querytweettime was "31692185296441344", the SearchSQL query was as follows:

```
SELECT RELEVANCE('2:3') AS REL, DOCNO, FT_CID
FROM TW12
WHERE ((FT_TEXT IS_ABOUT 'Australian Open Djokovic vs. Murray')
  AND (DOCDATE CONTAINS '28' WEIGHT 0|'29' WEIGHT 0|'30' WEIGHT 0|'31' WEIGHT 0))
ORDER BY REL DESC;
```

From this first-pass query, the top 100,000 matches (based on the relevance function) were retrieved (or all the matches if there fewer than 100,000).

As a second pass, a Java program was run on these (up to) 100,000 matches to remove the remaining tweets whose id was after the querytweettime.

As a third pass, another Java program was run to remove tweets which were not eligible because they were not listed as http status 200 in the previously mentioned organizer-provided "id-status.01-May-2012" file. Also in this pass, tweets starting with "RT" were discarded because "retweets" were also considered non-relevant by the task guidelines.

On the fourth and final pass, another Java program was run to just keep the first 10,000 remaining matches (or all the remaining matches if there were fewer than 10,000). There was also some other formatting for the official submission in this pass, e.g., remove the rank column that was not wanted this year.

There was also a post-hoc check that if a query ended up with fewer than the desired 10,000 matches, it was not the case that the first-pass heuristic of cutting off at 100,000 had filtered out some possible matches. (Otherwise we would have redone the run with a larger number in the first step.)

This run was labelled as using "future evidence" only because the inverse document frequency (idf) of each term was based on the term's frequency in the entire collection rather than just the tweets up to the given querytweettime.

2.3.2 HTTP Run - otM12ih

The submitted run 'otM12ih' was produced in the same way as the baseline run otM12i except that the word "http" was added to each query.

The motivation for doing so was that we had observed in the relevance assessments (qrels) of 2011, after crossing them with our download of the tweets, that a surprising 95% of the (downloaded) tweets judged highly relevant in at least one topic (503 of 530) included the string "http:", indicating that the highly relevant tweet referenced a web page via a url.

Also, 78% of the other (downloaded) tweets judged relevant in at least one topic (1753 of 2248) included the string "http:".

However, 52% of the remaining (downloaded) tweets that were judged in at least one topic (25205 of 48822) also included the string "http:" (these were tweets that were always judged non-relevant, though that they were judged likely means that some participant system last year gave a high rank to the tweet for at least one topic).

Overall, 18% of the entire (downloaded) collection of tweets contained the word "http" (2,419,519 of 13,551,650).

In our training experiments with last year's topics, adding the word "http" gave a modest boost to most of the evaluation metrics.

When "http" was added to the query, every query (even after all the time and eligibility filters) returned at least 10,000 tweets, unlike the baseline query.

Note that we did not ever download the web pages referred to by the tweets. (None of our runs used "external evidence" from outside the Tweets2011 corpus.)

2.3.3 No-IDF Run - otM12h

The submitted run 'otM12h' was produced in the same way as the "http" run otM12ih except that inverse document frequency (idf) was not used to weight the query terms. All query terms were given equal weight by specifying the SearchServer '2:5' relevance method instead of '2:3'.

Table 1: Mean Scores of Submitted Microblog Adhoc Search Runs

Run	GS30	FRS	S10	MRR	S1	P30	R-prec	GMAP	MAP
otM12h	0.925	0.854	53/59	0.660	31/59	0.375	0.286	0.128	0.231
otM12i	0.910	0.831	52/59	0.641	30/59	0.377	0.294	0.142	0.250
otM12ih	0.923	0.856	53/59	0.685	33/59	0.397	0.324	0.164	0.278
otM12ihe	0.891	0.811	48/59	0.662	33/59	0.411	0.326	0.158	0.299
(otM12ihe0)	0.855	0.768	46/59	0.617	31/59	0.411	0.317	0.117	0.288
[on judged]	GS30J	FRSJ	S10J	MRRJ	S1J	P30J	R-precJ	GMAPJ	MAPJ
otM12h	0.925	0.854	53/59	0.660	31/59	0.375	0.295	0.174	0.269
otM12i	0.910	0.831	52/59	0.642	30/59	0.377	0.302	0.186	0.281
otM12ih	0.923	0.856	53/59	0.685	33/59	0.397	0.327	0.214	0.307
otM12ihe	0.891	0.811	48/59	0.662	33/59	0.411	0.331	0.204	0.329
(otM12ihe0)	0.855	0.768	46/59	0.617	31/59	0.411	0.325	0.172	0.324
[highly rel]	HGS30	HFRS	HS10	HMRR	HS1	HP30	HP@R	HGMAP	HMAP
otM12h	0.793	0.695	41/56	0.494	21/56	0.225	0.229	0.083	0.200
otM12i	0.794	0.661	39/56	0.423	16/56	0.207	0.208	0.083	0.181
otM12ih	0.811	0.700	42/56	0.493	21/56	0.232	0.225	0.105	0.214
otM12ihe	0.787	0.676	39/56	0.486	21/56	0.243	0.248	0.094	0.238
(otM12ihe0)	0.759	0.645	38/56	0.448	19/56	0.247	0.246	0.067	0.227
[on judged h]									
otM12h	0.794	0.695	41/56	0.494	21/56	0.225	0.231	0.113	0.216
otM12i	0.794	0.661	39/56	0.424	16/56	0.207	0.209	0.106	0.192
otM12ih	0.811	0.700	42/56	0.493	21/56	0.232	0.226	0.129	0.225
otM12ihe	0.787	0.676	39/56	0.486	21/56	0.243	0.249	0.123	0.249
(otM12ihe0)	0.764	0.645	38/56	0.448	19/56	0.247	0.248	0.102	0.240

The task guidelines requested that at least one submitted run "not use any external or future source of evidence". This run satisfied that criteria.

2.3.4 Blind Feedback Run - otM12ihe

The submitted run 'otM12ihe' was a blind feedback run based 50% on otM12ih and 50% on an expansion query based on the first 30 pre-querytweettime tweets in the second-pass result of otM12ih.

Some notes:

Our feedback set excluded not just tweets after the querytweettime, but also the tweet of exactly the querytweettime (if it would have otherwise been in the top-30) because the guidelines originally disallowed this tweet. The organizers decided to allow this tweet on July 9, 2012 (the day before submissions were due) because the perl checker for submissions had accidentally been allowing it. We did not redo our feedback run at this time other than to not filter the querytweettime tweet from the final result.

Our feedback set included retweets and the tweets of non-200 status (likely from status 301 redirects). It's unclear whether keeping these in the feedback set was helpful or detrimental compared to filtering them out; we haven't done any experiments on this point.

Roughly speaking, the expansion query appended together the 30 tweets and filtered out the terms in more than 5% of the tweets in the full collection. It kept the remaining top-200 terms (based on stems actually) in a tf.idf calculation.

The result of the expansion query was saved in an (unsubmitted) run called otM12ihe0.

The final otM12ihe run was a fusion run based 50% on the base otM12ih run and 50% on the otM12ihe0 expansion run (based on the relevance() function score).

Table 2: Impact of Microblog Adhoc Search Techniques on Measures counting All Relevant

Expt	ΔGS30	95% Conf	vs.	3 Extreme Diffs (Topic)
e (ihe-ih)	−0.032	(−0.061,−0.004)	12-17-30	−0.47 (58), −0.41 (105), 0.15 (82)
h (ih-i)	0.013	(−0.003, 0.028)	12-5-42	0.23 (77), 0.18 (51), −0.19 (72)
i (ih-h)	−0.002	(−0.019, 0.014)	9-9-41	0.27 (72), −0.20 (53), −0.25 (51)
e0 (ihe0-ih)	−0.068	(−0.117,−0.019)	12-22-25	−1.00 (80), −0.63 (61), 0.15 (82)
	ΔFRS			
e (ihe-ih)	−0.044	(−0.095, 0.006)	12-17-30	−0.60 (80), −0.60 (61), 0.42 (82)
h (ih-i)	0.024	(−0.007, 0.055)	12-5-42	0.44 (77), 0.43 (107), −0.33 (72)
i (ih-h)	0.002	(−0.020, 0.023)	9-9-41	0.27 (99), −0.21 (54), −0.24 (51)
e0 (ihe0-ih)	−0.088	(−0.158,−0.018)	12-22-25	−1.00 (80), −0.96 (61), 0.42 (82)
	ΔMRR			
e (ihe-ih)	−0.023	(−0.122, 0.076)	12-17-30	−0.92 (61), −0.92 (80), 0.88 (82)
h (ih-i)	0.043	(−0.002, 0.089)	12-5-42	0.88 (59), 0.67 (99), −0.25 (110)
i (ih-h)	0.025	(−0.036, 0.086)	9-10-40	0.80 (99), 0.75 (96), −0.67 (98)
e0 (ihe0-ih)	−0.067	(−0.182, 0.047)	12-22-25	−1.00 (80), −0.98 (61), 0.88 (82)
	ΔP30			
e (ihe-ih)	0.014	(−0.006, 0.033)	23-20-16	0.23 (81), 0.20 (65), −0.13 (96)
h (ih-i)	0.020	(0.001, 0.039)	27-11-21	−0.23 (78), 0.17 (57), 0.17 (102)
i (ih-h)	0.022	(0.003, 0.041)	27-14-18	0.23 (110), 0.20 (83), −0.17 (79)
e0 (ihe0-ih)	0.014	(−0.009, 0.038)	23-23-13	0.30 (81), 0.23 (79), −0.17 (61)
	ΔR-prec			
e (ihe-ih)	0.003	(−0.017, 0.022)	25-19-15	0.29 (79), −0.17 (105), −0.25 (84)
h (ih-i)	0.030	(0.016, 0.044)	32-6-21	0.14 (79), 0.14 (102), −0.09 (78)
i (ih-h)	0.037	(0.015, 0.060)	34-15-10	0.29 (110), 0.26 (66), −0.16 (79)
e0 (ihe0-ih)	−0.007	(−0.030, 0.017)	22-26-11	0.31 (79), 0.19 (65), −0.25 (84)
	ΔGMAP'			
e (ihe-ih)	−0.003	(−0.019, 0.012)	36-22-1	−0.34 (61), −0.15 (80), 0.08 (72)
h (ih-i)	0.012	(0.004, 0.020)	47-10-2	0.15 (59), −0.07 (78), −0.10 (72)
i (ih-h)	0.021	(−0.001, 0.043)	44-15-0	−0.42 (63), 0.26 (110), 0.27 (72)
e0 (ihe0-ih)	−0.030	(−0.060, 0.001)	35-23-1	−0.51 (80), −0.49 (96), 0.09 (72)
	ΔMAP			
e (ihe-ih)	0.022	(0.004, 0.040)	36-22-1	0.24 (79), 0.17 (55), −0.23 (84)
h (ih-i)	0.028	(0.016, 0.040)	47-10-2	−0.13 (78), 0.11 (88), 0.11 (52)
i (ih-h)	0.047	(0.029, 0.064)	44-15-0	0.24 (66), 0.21 (71), −0.05 (79)
e0 (ihe0-ih)	0.011	(−0.014, 0.035)	35-23-1	0.29 (79), 0.22 (75), −0.26 (84)

This run was labelled as using "future evidence" only because the inverse document frequency (idf) of each term was based on the term's frequency in the entire collection rather than just the tweets up to the given querytweettime. No other future evidence was used; in particular, as already noted, the feedback set did not include any tweets from after the querytweettime.

3 Results

The track organizers released the relevance assessments (qrels) on August 31, 2012.

59 of the topics had at least one tweet judged relevant, averaging 106.5 relevant tweets per topic (ranging from as low as 1 relevant tweet for a topic to as high as 572).

Table 3: Impact of Microblog Adhoc Search Techniques on Measures counting just Highly Relevant

Expt	ΔHGS30	95% Conf	vs.	3 Extreme Diffs (Topic)
e (ihe-ih)	−0.024	(−0.070, 0.021)	17-19-20	−0.93 (61), −0.37 (58), 0.28 (101)
h (ih-i)	0.017	(−0.020, 0.054)	24-8-24	−0.62 (72), −0.30 (101), 0.60 (102)
i (ih-h)	0.018	(−0.027, 0.063)	12-13-31	0.93 (61), 0.60 (110), −0.37 (96)
e0 (ihe0-ih)	−0.053	(−0.115, 0.010)	19-23-14	−1.00 (80), −0.93 (61), 0.38 (59)
	ΔHFRS			
e (ihe-ih)	−0.024	(−0.081, 0.033)	17-18-21	−0.79 (61), −0.60 (80), 0.63 (82)
h (ih-i)	0.039	(0.006, 0.072)	22-8-26	0.46 (102), 0.44 (77), −0.34 (101)
i (ih-h)	0.005	(−0.040, 0.050)	10-13-33	0.79 (61), −0.43 (82), −0.64 (96)
e0 (ihe0-ih)	−0.055	(−0.126, 0.015)	18-23-15	−1.00 (80), −0.79 (61), 0.63 (82)
	ΔHMRR			
e (ihe-ih)	−0.007	(−0.094, 0.079)	17-18-21	0.93 (82), −0.83 (52), −0.92 (80)
h (ih-i)	0.070	(0.023, 0.116)	24-8-24	0.67 (71), 0.67 (99), −0.17 (109)
i (ih-h)	−0.001	(−0.043, 0.041)	12-14-30	0.80 (99), 0.25 (61), −0.67 (68)
e0 (ihe0-ih)	−0.045	(−0.154, 0.063)	19-23-14	−1.00 (80), −0.86 (75), 0.93 (82)
	ΔHP30			
e (ihe-ih)	0.011	(−0.006, 0.027)	19-16-21	0.17 (99), 0.17 (62), −0.13 (96)
h (ih-i)	0.025	(0.009, 0.041)	24-6-26	0.30 (66), 0.13 (96), −0.07 (101)
i (ih-h)	0.007	(−0.006, 0.020)	17-13-26	0.17 (110), 0.13 (83), −0.13 (79)
e0 (ihe0-ih)	0.015	(−0.005, 0.035)	19-19-18	0.23 (79), 0.20 (99), −0.13 (96)
	ΔHP@R			
e (ihe-ih)	0.022	(0.000, 0.045)	19-10-27	0.34 (79), 0.33 (109), −0.10 (64)
h (ih-i)	0.017	(0.003, 0.031)	19-6-31	0.20 (64), 0.13 (99), −0.12 (98)
i (ih-h)	−0.003	(−0.020, 0.013)	12-12-32	−0.22 (79), −0.17 (103), 0.15 (83)
e0 (ihe0-ih)	0.020	(−0.006, 0.047)	20-15-21	0.39 (79), 0.33 (109), −0.18 (94)
	ΔHGMAP'			
e (ihe-ih)	−0.009	(−0.036, 0.018)	32-23-1	−0.67 (61), −0.17 (80), 0.11 (81)
h (ih-i)	0.020	(0.008, 0.032)	49-6-1	0.26 (59), 0.13 (102), −0.11 (72)
i (ih-h)	0.020	(−0.010, 0.051)	36-19-1	0.46 (72), 0.39 (61), −0.44 (63)
e0 (ihe0-ih)	−0.037	(−0.077, 0.004)	28-27-1	−0.71 (96), −0.67 (61), 0.11 (81)
	ΔHMAP			
e (ihe-ih)	0.024	(0.002, 0.046)	32-23-1	0.27 (109), 0.25 (79), −0.12 (84)
h (ih-i)	0.033	(0.021, 0.045)	49-6-1	0.18 (56), 0.14 (66), −0.05 (72)
i (ih-h)	0.014	(0.003, 0.025)	36-19-1	0.13 (83), 0.13 (110), −0.08 (79)
e0 (ihe0-ih)	0.013	(−0.017, 0.042)	28-27-1	0.35 (109), 0.30 (79), −0.23 (91)

56 of the topics had at least one tweet judged highly relevant, averaging 45.9 highly relevant tweets per topic (ranging from 1 to 322).

Table 1 lists the mean scores of the previously described 4 submitted runs and 1 additional run. (A glossary with the definitions of the measures is in Section 4.1.) We see that the robust measures, such as FRS and GS30, find that the plain word+http run (otM12ih) has a higher score than the blind feedback runs (otM12ihe and otM12ihe0). However, the non-robust traditional TREC measures, such as MAP, R-prec and P30, favor the non-robust blind feedback runs. All of the measures agree that adding the word 'http' to the query increased the mean score (otM12ih vs. otM12i).

Tables 2 and 3 isolate the differences between runs in more detail as explained in Section 4.2.

3.1 Impact of Adding HTTP to Query

The 'h' lines (otM12ih score minus otM12i score) of Tables 2 and 3 show that adding "http" to the query produced a statistically significant increase in the mean scores of both traditional TREC measures (P30, R-prec, GMAP', MAP, HP30, HP@R, HGMAP', HMAP) and some robust measures (HFRS, HMRR).

3.2 Impact of Inverse Document Frequency

The 'i' lines (otM12ih score minus otM12h score) of Tables 2 and 3 show that using full-collection idf produced a statistically significant increase in the mean scores of some of the traditional TREC measures (P30, R-prec, MAP, HMAP), but none of the robust measures were able to discern a statistically significant impact.

3.3 Impact of Blind Feedback

The 'e' lines (otM12ihe score minus otM12ih score) of Tables 2 and 3 show that the blind feedback technique produced a statistically significant increase in the mean scores of some of the traditional TREC measures (MAP, HP@R, HMAP), but a statistically significant decrease in the mean score of one of the robust measures (GS30).

We have seen this result before not just in our own experiments but in 7 other groups' blind feedback experiments at the 2003 RIA Workshop [7].

More background on robust retrieval: The objective of robust retrieval [11] is to reduce the frequency of very poor results. The blind feedback technique is considered non-robust because of its tendency to "not help (and frequently hurt) the worst performing topics" [11]. Success@10 (S10) has been suggested as a "direct measure" of robustness, but it "has the drawback of being a very coarse measure" [11]. We have introduced less coarse variants such as "First Relevant Score" [9] and "Generalized Success@30" [6] and have found that they often can discern statistically significant negative impacts from the non-robust blind feedback technique [7].

4 Glossary

4.1 Retrieval Measures

This section states the definition of all of the retrieval measures of Table 1.

Robust measures — these measures are based just on the rank of the first relevant item retrieved:

- *Success@n* (S@n): For a topic, Success@n is 1 if a relevant item is retrieved in the first n rows, 0 otherwise. This paper lists Success@10 (S10) and Success@1 (S1) for all runs.
- *First Relevant Score* (FRS): For a topic, FRS is 1.08^{1-r} where r is the rank of the first row for which a relevant item is found, or zero if a relevant item was not found. This measure was introduced in [9]. The measure is also known as Generalized Success@10 (GS10) because it rounds to 1 for $r \leq 10$ and to 0 for $r > 10$. Intuitively, FRS is approximately the percentage of topics for which a relevant item is returned in the first 10 rows.
- *Generalized Success@30* (GS30): For a topic, GS30 is 1.024^{1-r} where r is the rank of the first row for which a relevant item is found, or zero if a relevant item was not found. Compared to FRS, GS30 further de-emphasizes small differences at the top of the result list.
- *Reciprocal Rank* (RR): For a topic, RR is $\frac{1}{r}$ where r is the rank of the first row for which a relevant item is found, or zero if a relevant item was not found. "Mean Reciprocal Rank" (MRR) is the mean of the reciprocal ranks over all the topics.

Traditional TREC ad hoc search measures — these measures place most of their weight on additional relevant items for a topic (after the first one):

- *Precision@n*: For a topic, "precision" is the percentage of retrieved items which are relevant. "Precision@n" is the precision after n items have been retrieved. This paper lists Precision@30 (P30) for all runs.
- *Average Precision* (AP): For a topic, AP is the average of the precision after each relevant item is retrieved (using zero as the precision for relevant items which are not retrieved). By convention, AP is based on the first 1000 retrieved items for the topic. The score ranges from 0.0 (no relevant items retrieved) to 1.0 (all relevant items retrieved at the top of the list). "Mean Average Precision" (MAP) is the mean of the average precision scores over all of the topics (i.e. all topics are weighted equally).
- *R-Precision* (R-prec): For a topic, R-precision is Precision@R where R is the number of relevant items for the topic. It is also equivalent to Recall@R where "recall" is the percentage of relevant items retrieved.
- *Geometric MAP* (GMAP): GMAP (introduced in [12]) is based on "Log Average Precision" which for a topic is the natural log of the max of 0.00001 and the average precision. GMAP is the exponential of the mean log average precision.
- *GMAP'*: We also define a linearized log average precision measure (denoted GMAP') which linearly maps the 'log average precision' values to the [0,1] interval. For statistical significance purposes, GMAP' gives the same results as GMAP, and it has advantages such as that the individual topic differences are in the familiar -1.0 to 1.0 range and are on the same scale as the mean.

If a J suffix is appended to the measure code, the measure is just evaluated using judged items (as if unjudged items were not retrieved). In particular, GS30J, FRSJ, S10J, MRRJ, S1J, P30J, R-precJ, GMAPJ and MAPJ are the same as GS30, FRS, S10, MRR, S1, P30, R-prec, GMAP and MAP (respectively) except that only judged items are considered.

If a H prefix is affixed to the measure code, the measure is just evaluated counting highly relevant items as relevant. In particular, HGS30, HFRS, HS10, HMRR, HS1, HP30, HP@R, HGMAP and HMAP are the same as GS30, FRS, S10, MRR, S1, P30, R-prec, GMAP and MAP (respectively) except that only highly relevant items are counted as relevant.

The measures were calculated using the l07_eval utility (which is available at http://trec.nist.gov/data/legal09.html). The scores may differ from trec_eval because l07_eval discards topics with no relevant documents (e.g., the H measures are averaged over just the 56 applicable topics, not 60) and l07_eval does no re-ordering of documents tied in rsv score.

4.2 Difference Tables

For the comparison tables (i.e., Tables 2 and 3), the columns are as follows:

- "Expt" specifies the experiment (the codes of the two runs being compared are listed, indicating first run minus second run).
- "Δ" is the difference of the mean scores of the two runs being compared (the column heading says for which retrieval measure).
- "95% Conf" is an approximate 95% confidence interval for the mean difference (calculated from plus/minus twice the standard error of the mean difference). If zero is not in the interval, the result is "statistically significant" (at the 5% level).
- "vs." is the number of topics on which the first run scored higher, lower and tied (respectively) compared to the second run. These numbers should always add to the number of topics.

- "3 Extreme Diffs (Topic)" lists 3 of the individual topic differences, each followed by the topic number in brackets. The first difference is the largest one of any topic (based on the absolute value). The third difference is the largest difference in the other direction (so the first and third differences give the *range* of differences observed in this experiment). The middle difference is the largest of the remaining differences (based on the absolute value).

5 Conclusions

In this paper, we measured the effectiveness of various experimental search techniques not just with traditional TREC ad hoc search measures such as Average Precision, R-precision and Precision@30, but also with robust measures based on just the rank of the first relevant item retrieved such as First Relevant Score and Generalized Success@30. We reported not just the mean scores of the experimental approaches but also the largest per-topic impacts of the techniques for several measures. Our experiments were conducted in the context of the Real-Time Adhoc Search Task of the TREC 2012 Microblog Track which investigated the effectiveness of ad hoc search of a collection of more than 10 million tweets. For the experimental technique of favoring tweets with urls, we found that both the traditional and robust measures indicated statistically significant increases in the mean score. However, for an experimental blind feedback technique, a technique known to be non-robust as it typically makes poor results even worse, the traditional Average Precision measure indicated a statistically significant increase in the mean score, but some of the measures just based on the rank of the first relevant item successfully discerned a statistically significant decrease in the mean score from the non-robust technique. We continue to advocate that researchers investigate the impacts of search techniques on the rank of the first relevant item.

References

[1] Cross-Language Evaluation Forum web site. `http://www.clef-campaign.org/`

[2] Andrew Hodgson. Converting the Fulcrum Search Engine to Unicode. Sixteenth International Unicode Conference, 2000.

[3] NTCIR (NII-Test Collection for IR) Home Page. `http://research.nii.ac.jp/~ntcadm/index-en.html`

[4] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, M. Gatford. Okapi at TREC-3. Proceedings of TREC-3, 1995.

[5] Text REtrieval Conference (TREC) Home Page. `http://trec.nist.gov/`

[6] Stephen Tomlinson. Comparing the Robustness of Expansion Techniques and Retrieval Measures. Working Notes for the CLEF 2006 Workshop. Revised in LNCS 4730.

[7] Stephen Tomlinson. Early Precision Measures: Implications from the Downside of Blind Feedback. *SIGIR 2006*, pp. 705-706.

[8] Stephen Tomlinson. Enterprise, QA, Robust and Terabyte Experiments with Hummingbird SearchServer™ at TREC 2005. Proceedings of TREC 2005.

[9] Stephen Tomlinson. European Ad Hoc Retrieval Experiments with Hummingbird SearchServer™ at CLEF 2005. Working Notes for the CLEF 2005 Workshop.

[10] Stephen Tomlinson. Learning Task Experiments in the TREC 2011 Legal Track. Proceedings of TREC 2011.

[11] Ellen M. Voorhees. Overview of the TREC 2003 Robust Retrieval Track. Proceedings of TREC 2003.

[12] Ellen M. Voorhees. Overview of the TREC 2004 Robust Retrieval Track. Proceedings of TREC 2004.